espol Facultad de Ciencias Naturales y Matemáticas

Luis Campuzano
luidcamp@espol.edu.ec

Hopkins Herbas
hherbas@espol.edu.ec

Clasificación de pacientes escolares con posible trastorno de déficit de atención e hiperactividad (TDAH) mediante técnicas de Machine Learning

PROBLEMA

El TDAH es un trastorno que incluye problemas de concentración, dificultad para controlar conductas impulsivas o exceso de actividad. Afecta comúnmente a los niños de hasta doce años. A pesar de que no existe una cura, se puede tratar oportunamente a aquellos que lo padecen. En el Ecuador hay 2.6 millones de infantes entre 5 y 12 años, que es el rango de edad para prevenir el TDAH. Por esta razón, es necesario realizar un diagnóstico a fin de detectar a aquellos que requieren ayuda inmediata. Sin embargo, la escasez de expertos en el área impide diagnosticar a todos los niños de forma rápida y eficiente.

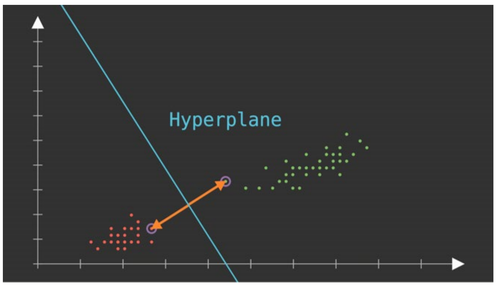
OBJETIVO GENERAL

Construir un algoritmo que clasifique de manera óptima a niños de primaria según la categoría de trastorno de déficit de atención e hiperactividad mediante técnicas de Machine Learning.

PROPUESTA

Un proceso de diagnóstico de TDAH requiere que los pacientes escolares, sus padres y sus profesores realicen un Test psicológico, como el propuesto por Conners y Brief. Luego, a través de un barómetro, se dictamina si el paciente tiene o no TDAH. Se propone agilizar los resultados de los cuestionarios mediante técnicas clasificación y predicción de los modelos de Machine Learning (ML).

Para clasificar a los niños que fueron evaluados con los Tests de Conners y Brief se estudiaron los algoritmos K- medias, K-vecinos más cercanos (K-NN) y Máquinas de vectores de soporte (SVM), siendo el más destacado el SVM. El algoritmo de ML se basa en la idea matemática de un hiperplano que separa las observaciones en dos grupos. El objetivo es obtener el separador óptimo para los datos. Básicamente minimiza la siguiente función objetivo:



$$J(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\{\boldsymbol{w}^T\phi(\boldsymbol{x}_n) - t_n\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

En la que w es el vector cuyas entradas se estiman al minimizar $J, \phi(\boldsymbol{x}_n)$ es la transformación del método del kernel evaluada en la observación $\boldsymbol{x}_n$, t_n es la etiqueta a la que pertenece el punto $\boldsymbol{x}_n$ y $\lambda \geq 0$ es un parámetro de regularización.

Este trabajo permite acelerar el proceso de diagnóstico de TDAH por parte de los especialistas, pues los modelos seleccionados y los ordenadores son mucho más rápidos clasificando que las personas, además, no se requiere de un costoso barómetro.

RESULTADOS

Para validar los resultados obtenidos se usó una matriz de confusión que muestra la cantidad de aciertos del algoritmo de Machine Learning. En la Figura 1 se observa que la matriz de confusión obtenida no es diagonal, ya que el algoritmo clasificó a un paciente como Desarrollo Típico cuando la realidad es que el paciente pertenece al grupo Combinado, mientras que de cuatro pacientes que pertenecen al grupo Hiperactivo, dos fueron clasificados como Desarrollo Típico.

Caso real / Predicción	Combinado	Hiperactivo	Inatento	Desarrollo Típico
Combinado	1	0	0	0
Hiperactivo	0	2	0	0
Inatento	0	0	0	0
Desarrollo Típico	1	2	0	34

Porcentajes / Grupos	Grupo de Entrenamiento (%)	Grupo de Prueba (%)
Segundo-Hombres-Conners	0	7.5
Segundo-Mujeres-Conners	2.38	10
Cuarto-Hombres-Conners	2.79	11.32
Cuarto-Mujeres-Conners	2.47	10.34
Sexto-Hombres-Conners	1.37	9.61
Sexto-Mujeres-Conners	0.44	16.66
Segundo-Hombres-Brief	0.62	5.12
Segundo-Mujeres-Brief	2.72	11.42
Cuarto-Hombres-Brief	6.1	13.46
Cuarto-Mujeres-Brief	1.69	25.42
Sexto-Hombres-Brief	0.55	6.66
Sexto-Mujeres-Brief	7.81	15.55

El porcentaje de eficacia en un algoritmo es el indicativo de qué tan bien trabaja el modelo para solucionar el problema en cuestión. Se calculó el porcentaje de eficacia para la data de entrenamiento, usando el 80% de los datos; y para la data de prueba, con el resto de los datos. La tabla a la izquierda muestra los resultados obtenidos.

Los porcentajes de error con los datos de entrenamiento son bajos indicativo que la investigación realizada es bastante buena.

CONCLUSIONES

- Con base en los resultados obtenidos, se evidencia que se logró obtener el algoritmo que mejor clasifica a los niños de primaria, permitiendo a los psicólogos tener un soporte para diagnosticar más rápido a los pacientes escolares, sin la necesidad de acudir a barómetros especializados costosos.
- Los modelos K-medias, K vecinos más cercanos y Máquinas de vectores de soporte tienen sus respectivas ventajas y desventajas. Con la matriz de confusión se pudo percatar que, de acuerdo a los datos usados, el SVM ofrece mejor rendimiento.